Statistical Approaches to Use a Model Organism for Regulatory Sequences Annotation of Newly Sequenced Species

Pietro Liò[1]*, Claudia Angelini[2], Italia De Feis[2], Viet-Anh Nguyen[1]

1 Computer Laboratory, University of Cambridge, Cambridge, United Kingdom, **2** Istituto per le Applicazioni del Calcolo "Mauro Picone" (CNR), Napoli, Italy

Abstract

A major goal of bioinformatics is the characterization of transcription factors and the transcriptional programs they regulate. Given the speed of genome sequencing, we would like to quickly annotate regulatory sequences in newly-sequenced genomes. In such cases, it would be helpful to predict sequence motifs by using experimental data from closely related model organism. Here we present a general algorithm that allow to identify transcription factor binding sites in one newly sequenced species by performing Bayesian regression on the annotated species. First we set the rationale of our method by applying it within the same species, then we extend it to use data available in closely related species. Finally, we generalise the method to handle the case when a certain number of experiments, from several species close to the species on which to make inference, are available. In order to show the performance of the method, we analyse three functionally related networks in the *Ascomycota*. Two gene network case studies are related to the G2/M phase of the *Ascomycota* cell cycle; the third is related to morphogenesis. We also compared the method with MatrixReduce and discuss other types of validation and tests. The first network is well known and provides a biological validation test of the method. The two cell cycle case studies, where the gene network size is conserved, demonstrate an effective utility in annotating new species sequences using all the available replicas from model species. The third case, where the gene network size varies among species, shows that the combination of information is less powerful but is still informative. Our methodology is quite general and could be extended to integrate other high-throughput data from model organisms.

Citation: Liò P, Angelini C, De Feis I, Nguyen V-A (2012) Statistical Approaches to Use a Model Organism for Regulatory Sequences Annotation of Newly Sequenced Species. PLoS ONE 7(9): e42489. doi:10.1371/journal.pone.0042489

Editor: Chuhsing Kate Hsiao, National Taiwan University, Taiwan

Received November 30, 2011; **Accepted** July 9, 2012; **Published** September 11, 2012

Funding: These authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: pl219@cam.ac.uk

Introduction

One of the most important and time consuming step in annotating a new genome is the identification of the transcription factor binding sites [1,2]. An important reason for such difficulty is their fast evolution with respect to coding regions, which limits the use of model organisms annotation [3]. Recently, due to the direct sequencing of all DNA fragments from ChIP assays, ChIP-Seq has become the best technology for genome-wide mapping of protein-DNA interactions [4].

An important class of binding site identification methods is based on the assumption that co-expressed groups of genes often share regulatory elements, which mediate the co-expression; interesting counter examples are described in [5]. A two-step approach is most commonly used. In the first step, the co-expressed groups of genes need to be determined, typically from gene-expression data. A clustering procedure is performed to partition the genes into groups believed to be co-regulated, based on expression profile similarity. In the second step, a motif discovery tool is applied to search for abundant sequence patterns in the promoters (or $3'$-UTRs) of each group that may represent the binding sites of transcription factors that regulate the corresponding genes. In [6] the authors applied linear regression with stepwise selection on a list of candidate motifs obtained using MDScan (see [7]) which is an algorithm that makes use of word-enumeration and position-specific probability matrix updating techniques. The candidate motifs were scored in terms of number of sites and degree of matching with each gene. Inspired by Liu's work, our group has explored the performances of algorithms based on Bayesian variable selection techniques showing that they can be more effective than stepwise regression [8],[9],[10]. In particular, in [10] and [8] we described a Bayesian variable selection model to take into account the different and multiple information sources available, to pool together results of several experiments and to allow the users to select the motifs that best explain and predict the changes in expression level in a group of co-regulated genes. When experiments are costly, particularly in high throughput biology, replicates come often in a minimum number to assure statistical reliability for disseminating and publishing results. In some cases, recently diverged species might retain similarities in gene expression. These considerations suggest that, in absence of experimental replicates, or even in addition to these, statistical support to experimental evidences may also be obtained by analysing model organisms that are phylogenetic close variants of the species under examination. The effective exploitation of annotated species richness is hampered by the lack of a robust theoretical statistical framework to combine and contrast the knowledge from replicas and from the model organisms nearby species. Here we describe a new systematic genome-wide statistical approach for identifying putative transcription factor binding sites from over-represented DNA sequence elements, or motifs, of newly sequenced species, by regressing gene expression data of nearby model species. The phylogenetic relationship between the species, using coding regions, is carried out for the sole purpose of identifying those model species that are enough close to potentially share similarity in gene expression and motifs. Then we use Bayesian variable selection to combine the information of the DNA sequences of the species under analysis with the genome expression information of other sufficiently close species, from which several experimental results are available. Pooling information across studies can help to accurately identify the true target genes, as pointed out in [11], allowing both to share the final cost of the analysis and to use already available data which are contained in classical repositories. The paper is organized as follows. First we set the rationale of our method by applying it within the same species, then we extend it to use data available in closely related species. With respect to previous publications, here we present a general algorithm to identify transcription factor binding sites in one species and perform Bayesian regression on the annotated species. Our generalisation could handle the case when a certain number of experiments from several species closed to the species on which to make inference are available. We also introduce an internal testing analysis and we investigate three different networks; then we compare results with those obtained using MatrixReduce, one of the best performing and used algorithm in the field [12]. Finally we discuss the findings on the three networks and we describe the statistical methodology in the Section Methodology.

Materials and Methods

Algorithm

The algorithm consists of three major stages: sequence processing, candidate motif selection, and motif detection. In the following subsections we describe the specific steps we carried out.

Sequence Preprocessing Steps

1) Select a group of co-regulated genes in a well-annotated species and collect related microarray expression experiments. In our study we considered three case studies with different model organisms: the septation transcriptional network in *S. Pombe*; the cytokinesis transcriptional network and RAM signalling network, both in *C.albicans*.

2) Determine the nearby species using phylogenetic properties of the gene set selected in the previous step. Phylogenetic analysis can be conducted with several methodologies. In the first case study, we assessed the distance among fungi species based on all Ace2p related protein trees using the JTT amino acid substitution model [13]. The choice was due to the fact that these proteins are globular cytoplasmic proteins. Likelihood maximization and maximum likelihood parameter estimation were performed by numerical optimization routines using a single replacement matrix for all sites. Based on phylogenetic similarity we selected *S. Japonicus* and *S. Octosporus* as nearby species of *S. Pombe*. In the second case study, we generated all RAM related trees and picked three candida genomes *C. Tropicalis*, *C. Dublinensis* and *C. Parapsilosis* as nearby species of *C. Albicans*. Note that the latter two species seem to be more distant from the first two species.

3) Choose a set of biologically independent genes for each model species (*S. Pombe*/*C. Albicans*) from the pool of remaining genes (those not selected in step (1)). This step is motivated from the fact that extensive comparative genomic analysis has revealed that all the eukaryotic genomes contain families of duplicated genes which have recently diverged. In many cases these families have retained large part of the upstream regulatory sequences. In particular the residues of whole genome duplications have been identified in different yeast strains [14] as well as in other species. The redundancy of yeast genome suggests us to select a meaningful non redundant ensemble of genes that contains all the relevant statistical characteristics of the genome and therefore will play the role of control genes in step (6).

To this purpose, we performed a phylogenetic analysis of the gene pool using standard maximum likelihood techniques. The analysis allowed us to identify subset of genes with very large sequence similarities and therefore may derive from a common ancestor. In all cases we also used GO slim annotations [15] as a guidance for genes likely functionality. Aimed to get a fair share of representatives from all functional and phylogenetic gene sets, we randomly sampled genes from each set such that the number of samples was proportionally to the set size. We ended up with approximately 500 background genes for case study 1, and 600 genes for case study 2 and 3.

4) Identify homologous genes in nearby species (case study 1: *S. Japonicus* and *S. Octosporus*; case study 2 and 3: *C. Tropicalis*, *C. Dublinensis* and *C. Parapsilosis*) for both the co-regulated and background sets. We used a recent homology map [16] to facilitate this step.

5) Extract upstream DNA sequences (1000 base pairs length) for each species, shorten them in case of overlapping with adjacent ORFs. For genes with negative orientation, we considered the reverse complement of the sequences. Note that motif finding algorithms are sensitive to noise, which increases with the size of upstream sequences examined, moreover the vast majority of the yeast regulator sites from the TRANSFAC database are located within 800 bp from the translation start sites [17].

Selection of candidate motifs

6) Generate candidate motifs enriched in promoter regions of co-regulated genes and compute their matching scores for each gene. We used a modified version of the software MDSCAN [7] to search for nucleotide patterns which appears in the upstream sequences of the genes of interest for each species. To remove repeated segments that might confuse the motif discovery process, we preprocessed the upstream sequences of the interested network genes using RepeatMasker [18]. The matching score between a candidate motif m and a given gene sequence g was calculated as in [6]:

$$X_{mg} = \log_2\left[\sum_{x \in Q_{wg}} \Pr(x\,\mathrm{from}\,\theta_m)/\Pr(x\,\mathrm{from}\,\theta_0)\right]$$

where Q_{wg} is the set of all w-mer in the upstream region of gene g. θ_m is the probability matrix of motif m of width w, θ_0 is the transition probability matrix for the background model, computed using a Markov chain of the sixth order (Liu's original algorithm permits only Markov chain of the third order) from the upstream regions of all the species of interest. We examined nucleotide patterns of length 5 to 12 bp and scored up to 30 distinct candidates for each width in all case studies.

Variable Selection and Inference

7) Identify likely regulatory motifs among the candidate sets obtained in the previous step. We used an extended version of the Bayesian variable selection [9] that handles the case of multiple

experiments [19]. The idea is to search for the set of motifs that provide the best fit when regressing nucleotide pattern matching scores (X) to the set of gene expression levels ($Y = \{Y^e\}_{e=1}^{E}$). $Y^e \in \mathbb{R}^{G \times K_e}$ contains the expression data from experiment e of G genes of the annotated species over K_e technical replicates. Pattern scores X is of size $G \times M$ where M is the number of candidate motifs, evaluated on the nearby species. We assume the following model

$$y_{gek}|\mu_{ge}, \sigma_e^2 \sim \mathrm{N}\left(\mu_{ge}, \sigma_e^2\right) \\ g=1,\ldots,G;\; k=1,\ldots,K_e;\; e=1,\ldots,E \tag{1}$$

where y_{gek} represents the observed gene expression value of the gene g in the k^{th} replicate of the e^{th} experiment and μ_{ge} is the underlying transcriptional activity level of gene g under the experimental condition e. A binary latent vector, γ of dimension M, is introduced to indicate the inclusion of variables in the model; γ_m takes on the value of 1 if the m^{th} variable (motif) is included and 0 otherwise. Let $m_\gamma = \sum_{i=1}^{M} \gamma_i$ be the total number of motifs included. The true gene expression value μ_{ge} is connected to a specific subset of the M candidate motifs identified by the latent vector γ by the following relation

$$\mu_{ge}|\gamma = \sum_{\{m:\gamma_m=1\}} x_{gm}\beta_{me} = \mathbf{x}_{g.(\gamma)}\,\beta_{e(\gamma)},$$

where $\mathrm{x}_{g.(\gamma)}$ is the row vector of matching scores for all included motifs against gene g. We specified the following priors for the regression coefficients, the experiment variance, and the latent indicator:

$$\beta_{e(\gamma)}|\sigma_e^2,\gamma \sim \mathrm{N}\left(0,\sigma_e^2 H_{(\gamma)}\right), \quad \sigma_e^2 \sim \mathrm{IG}(v,S), \\ p(\gamma) = \Pi_{j=1}^{M}\theta^{\gamma_j}(1-\theta)^{1-\gamma_j}, \tag{2}$$

where prior parameters $H_{(\gamma)}$,v and S were assessed by a sensitivity analysis and $\theta = m_{\text{prior}}/M$ with m_{prior} as the number of covariates expected *a priori* to be included in the model. For all case studies we chose $H_{(\gamma)} = c\left[\operatorname{diag}(X'X)^{+}\right]_{(\gamma)}$, where X is the score matrix obtained in step (6) and c equals to the variation of the regression coefficients of the full model averaged over the experiments. We set weak prior knowledge choosing $v=3$, which is the smallest integer such that the expected noise level of an experiment $\mathbb{E}(\sigma_e^2) = v/(v-2)S$ exists. The scaling value S is equal to data variation averaged over all experiments. For case study 1, whose data are from time-course experiments, y_{gek} represents the average value of gene expression levels measured in the interval when the ENG1 genes show their common activity peak, approximately 30–90 minutes. The model specified in equation (1) could be rewritten as

$$\mathrm{y}_{.k}^{e} \sim \mathrm{N_G}\left(X_{(\gamma)}\beta_{e(\gamma)},\sigma_e^2 I\right)$$

where $\mathrm{y}_{.k}^{e} = (y_{1k}^{e},\ldots,y_{Gk}^{e})'$, $k=1,\ldots,K_e$ and $e=1,\ldots,E$. Without loss of generality we further assume that the columns of X and $\mathrm{y}_{.k}^{e}$ are mean-centered.

Having set the prior distributions, a Bayesian analysis proceeds by updating the prior beliefs with information that comes from the data. The posterior distribution of the latent indicator vector γ given the data, i.e., $f(\gamma|X,Y^1,\ldots,Y^E)$, can be obtained:

$$f(\gamma|X,Y^1,\ldots,Y^E) \propto \\ \Pi_{e=1}^{E} a_e \frac{|H_{(\gamma)}|^{-1/2}|K_{(\gamma)}^{e}|^{-1/2}}{\left(c_e - M_e'\left(K_{(\gamma)}^{e}\right)^{-1} M_e + S\right)^{(GK_e+v)/2}} \tag{3}$$

with

$$K_{(\gamma)}^{e} = \left(K_e X_{(\gamma)}{}' X_{(\gamma)} + H_{(\gamma)}^{-1}\right)$$

$$c_e = \sum_{k=1}^{K_e}\left(\mathrm{y}_{.k}^{e}\right)\left(\mathrm{y}_{.k}^{e}\right)$$

$$M_e = X_{(\gamma)}{}'\left(\sum_{k=1}^{K_e}\mathrm{y}_{.k}^{e}\right)$$

$$a_e = (1/2\pi)^{GK_e/2} S^{v/2}\left[\Gamma((GK_e+v)/2)2^{(GK_e+v)/2}\right] / \left[\Gamma(v/2)2^{v/2}\right].$$

The model (1)–(2) could be generalized in order to handle the presence of missing data which are typically encountered when analyzing real data experiment. For each fixed experiment $e=1,\ldots,E$ and each fixed technical replicate $k=1,\ldots,K_e$, let G_{es} be the number of genes with expression levels measured on the array. In this case the posterior becomes

$$f(\gamma|X_{ek},\mathrm{y}_{.k}^{e},k=1,\ldots,K_e;e=1,\ldots,E) \propto \\ \Pi_{e=1}^{E} a_e \frac{|H_{(\gamma)}|^{-1/2}|K_{(\gamma)}^{e}|^{-1/2}}{\left(c_e - M_e'\left(K_{(\gamma)}^{e}\right)^{-1} M_e + S\right)^{(G_e+v)/2}} \tag{4}$$

with

$$G_e = \sum_{k=1}^{K_e} G_{ek}$$

$$K_{(\gamma)}^{e} = \left(\sum_{k=1}^{K_e} X_{ek(\gamma)}' X_{ek(\gamma)} + H_{(\gamma)}^{-1}\right), \quad X_{ek} \quad \text{of} \quad \text{dimension} \quad G_{ek} \times M$$

$$c_e = \sum_{k=1}^{K_e}\left(\mathrm{y}_{.k}^{e}\right)\left(\mathrm{y}_{.k}^{e}\right), \quad \mathrm{y}_{.k}^{e} \quad \text{of} \quad \text{dimension} \quad G_{ek} \times 1$$

$$M_e = \sum_{k=1}^{K_e} X_{ek(\gamma)}' \mathrm{y}_{.k}^{e}$$

$$a_e = (1/2\pi)^{G_e/2} S^{v/2}\left[\Gamma((G_e+v)/2)2^{(G_e+v)/2}\right] / \left[\Gamma(v/2)2^{v/2}\right].$$

Our interest is to maximize the posterior probability in equations (3) and (4). Since the relative high dimensionality of our vector space (approximately $M=150$ for our case studies) makes comprehensive evaluation of posterior probabilities impossible, we employed a sampling procedure based on stochastic search Markov Chain Monte Carlo (MCMC) technique to identify realizations of γ with huge posterior probabilities.

8) Run multiple parallel MCMC chains of significant length for each species. In each run, the algorithm visits a sequence of models that differ successively in one or two variables. At each iteration, a candidate model, represented by γ^{new}, is generated by randomly choosing one of these two transition moves:

(i) Add or delete one variable from γ^{old}.
(ii) Swap the inclusion status of two variables in γ^{old}.

The proposed γ^{new} is accepted with a probability that depends on the ratio of the relative posterior probabilities of the new versus the previously visited models:

$$\min\left\{\frac{f(\gamma^{\text{new}}|X,Y^1,\ldots,Y^E)}{f(\gamma^{\text{old}}|X,Y^1,\ldots,Y^E)},1\right\}, \tag{5}$$

which leads to the retention of the more probable set of patterns. An analogous formula is obtained considering the posterior probability given by formula (4).

The stochastic search results in a list of visited sets (i. e. combination of candidate motifs) and the corresponding relative posterior probabilities, then the selection of few "best motifs" can be done either using the global MAP principle or by selecting the covariates on the basis of their marginal probability to be included. The marginal posterior probability of inclusion for a single motif j, $P(\gamma_j=1|X,Y^1,\ldots,Y^E)$, can be computed by averaging out the posterior probabilities of the acquired samples:

$$\begin{aligned}
f(\gamma_j=1|X,Y^1,\ldots,Y^E) &= \int f\left(\gamma_j=1,\gamma_{(-j)}|X,Y^1,\ldots,Y^E\right)d\gamma_{(-j)}\\
&\propto \int f\left(Y^1,\ldots,Y^E|X,\gamma_j=1,\gamma_{(-j)}\right)\cdot f(\gamma)d\gamma_{(-j)} \qquad (6)\\
&\approx \sum_t f\left(Y^1,\ldots,Y^E|X,\gamma_j=1,\gamma^{(t)}_{(-j)}\right)\cdot f\left(\gamma_j=1,\gamma^{(t)}_{(-j)}\right),
\end{aligned}$$

where $\gamma^{(t)}_{(-j)}$ is the vector γ at the t^{th} iteration without the j^{th} motif.

In all three case studies, we ran 10 parallel chains of 100,000 iterations each. We computed the normalized posterior probabilities for each distinct visited set of motifs and the marginal probabilities for the inclusion of single nucleotide patterns.

Robustness analysis

To investigate the effect of sparsity setting on variable selection, we ran steps 7)–8) with various values for m_{prior}, the a-priori expected number of motifs included in the model. In particular we examined $m_{\text{prior}}=1,3,5$ for all three case studies. These values were chosen due to the knowledge that fungi are simple organisms and their regulation mechanisms are based on relatively few motifs.

To study the robustness of the proposed framework with respect to the choices of both single experiment and control gene sets, we repeated steps 7)–8) with different subsets of control genes in combination with the leave-one-out cross validation strategy over all experiments. In our case studies we randomly sampled 8 different subsets of 200 genes out of 500–600 background genes in total.

Internal testing analysis

The procedure described above is based on the implicit assumption that the expression levels of the co-regulated genes observed in a microarray experiment Y^e of one species are positively correlated to those of the homologous gene set in the species of interest. In the case experimental data are not available for the latter species, we could validate such assumption using a third species of larger phylogenetic distance than the species under studies. Let Z^e be the expression data of the third species, we could compute the correlation coefficient of the co-regulated gene set in Y^e and Z^e. If the coefficient is significantly high (i.e. close to 1), we deduce that the assumption is likely to be satisfied. In this particular context for the first dataset we computed the correlation coefficients between *S. Pombe* and *S. Cerevisiae* in order to justify the comparison of the networks of *S. Pombe*, *S. Octosporus* and *S. Japonicus*. This approach requires a good degree of agreement from several species as criterion of trust of the solution.

Further generalisation

Our method can be generalized to handle the case when multiple experiments from different species close (phylogeneticaly) to the species under investigation are available. A straightforward solution is to run the described model independently for each species and pool out proposed models by their marginal probabilities. An alternative proposal is to incorporate all information available into a single model as follows.

Let $y_{sge_sk_{e_s}}$ be the observed gene expression value of the gene g in replicate k_{e_s} of experiment e_s from species s, with $s=1,\ldots,S$; $g=1,\ldots,G$, $e_s=1,\ldots,E_s$, and $k_{e_s}=1,\ldots,K_{e_s}$. As before, we assume expression values follow a normal distribution,

$$y_{sge_sk_{e_s}}|\mu_{sge_s},\sigma^2_{e_s}\sim \mathrm{N}\left(\mu_{sge_s},\sigma^2_{e_s}d_s\right),$$

where d_s is proportional to the distance between species s and the species of interest and is estimated from the phylogenetic tree. The distances are normalized such that $0\leq d_s\leq 1$ and $\sum_{s=1}^{S} d_s=1$. Let $\underline{\mu}_s$ be a matrix of dimension $G\times E_s$, which is defined as

$$\underline{\mu}_s=(\underline{\mu}_1,\ldots,\underline{\mu}_{E_s})$$

where $\underline{\mu}_{e_s}$ is a column vector of length G and represents the genes expression values in the e_s^{th} experiment for species s. We further assume a normal matrix variate distribution on $\underline{\mu}_{e_s}$ as follows:

$$\underline{\mu}_s|B_s,\Omega_s\sim \mathrm{N}(XB_s,I_G,d_s\Omega_s),$$

where B_s is the coefficient matrix of dimension $M\times E_s$, $\Omega_s=\mathrm{diag}(\sigma^2_{e_s})$ is the covariance matrix of dimension $E_s\times E_s$ over experiments, and I_G is the identity matrix of dimension $G\times G$ over genes. We assume varying noise levels over experiments, but fixed global noise for biological systems. Conjugate priors are employed for the coefficients B_s and the covariance matrix Ω_s:

$$B_s|\Omega_s\sim \mathrm{N}\left(0,H_\gamma,d_s\Omega_s\right) \quad \text{and} \quad \sigma^2_{e_s}\sim \mathrm{IG}(v_s,T_s),$$

where H_γ is the covariance matrix for the motifs of dimension $M_\gamma\times M_\gamma$, assumed to be known a priori, and v_s and T_s depends on the species s. By integrating out B_s and $\sigma^2_{e_s}$ and computing the posterior $f(\gamma|X,Y)$, inference can be performed similarly to the previously described procedure.

Results and Discussion

Case Study 1: septation transcriptional network in fission yeast clade

One of the key biological processes in the cell is the cytokinesis during which daughter cells separate and form two independent entities. In many unicellular fungi such as the fission yeast *Schizosaccharomyces Pombe*, a contractile actomyosin ring (CAR) generates a cell cleavage and the newly synthesized membrane is inserted at the division site. *S. Pombe* cells then divide by medial fission through the contraction of an actomyosin ring and the deposition of a multilayered division septum that must be cleaved to release the two daughter cells. Seven genes (adg1, adg2, adg3, cfh4, agn1, eng1, and mid2) whose expression is induced by the transcription factor Ace2p have been identified. Their transcription levels vary during the cell cycle, while maximum transcription are observed during septation [20,21]. The division septum has a threelayer structure, with a central primary septum (mainly composed of linear β-1,3-glucan) surrounded on both sides by two secondary septa (composed of β-1,6- branched β-1,3-glucan and β-1,6-glucan). The primary septum is synthesized through action of the Cps1/Bgs1 glucan synthase, while Bgs4 is involved in the assembly of secondary septa. Daughter cell separation requires an enzymatic process that controls the degradation of the components of the primary septum and the surrounding cell wall. To date, the two main enzymatic activities identified are exerted by the endo-β-1,3-glucanase Eng1, which is responsible for primary septum hydrolysis, and the endo-α-1,3-glucanase Agn1, which is necessary for the erosion of the cylinder of the cell wall surrounding the septum. The pattern of activation of Eng1 involves Sep1p, a protein of the conserved forkhead family. This protein targets a gene which encodes the transcription factor, Ace2p. Two of the Ace2p target genes encode proteins with known roles in cell separation: the β-glucanase Eng1p, that degrades the primary division septum between the new ends of daughter cells, and the α-glucanase Agn1p, that hydrolyses the old cell wall surrounding the septum leading to full separation of daughter cells [22]. Cells that constitutively overexpress Ace2 become round and show high transcript levels for both Eng1 and Agn1. The round shape of these cells could reflect a weakening of cell wall material that is not associated with the division septum, caused by an overproduction of glucanases [23,24]. Both [25] and [22] have found the motifs CC(T/A)CG(T/C)TCC, and (A/T)ACC(T/A)CGC(T/A). Interestingly, the consensus site for Ace2 (CCAGCC) is reminiscent of the core of New 3v (CCACGC), suggesting that an unknown Ace2-like factor could be involved.

We applied the Bayesian variable selection framework (as described in the Materials and Methods section) to detect binding motifs that regulate the network through various cell cycle phases. We obtained expression data from experiment elutriation A described in [25] and experiments elutriation 1, elutriation 2, elutriation 3 and cdc25 block release 1 described in [22]. The experiments explore the transcriptional activity of the fission yeast *S. Pombe* as a function of time in cells synchronized by different approaches: centrifugal elutriation and the use of temperature sensitive cell cycle mutants. All these experiments have no technical replicates. The upstream sequences for *S. Pombe*, *S. Japonicus* and *S. Octosporus* were obtained from the MIT Broad Schizosaccharomyces database [26].

Motifs detected with corresponding marginal probabilities larger than 0.5 are shown in Table 1. For the sake of space we present only the results obtained using all the 5 experiments for $m_{\text{prior}} = 1$. Marginal probabilities were averaged over 8 subsets of control genes. We obtained both confirmation of known results and new findings (motifs) which have high marginal probability values. We note that long patterns were selected more often than short ones. This could be explained by the limited ability to reject associations among nearby DNA bases of the background model. Eukaryotic DNA is highly heterogeneous, patchy and repetitious [27] and currently used genome background models cannot adequately take into account the variations in base association. We also observe that given more replicates or data from more species, the marginal probabilities become much higher (about three-fold) than those obtained using single replicate and one species, see [8]. A good understanding of how genes involved in this network differ in nearby species is provided by phylogenetic inference. Figure 1 shows the maximum likelihood phylogenetic tree obtained using ENG1 protein from a large number of fungi species (see legend), including *S. Japonicus*, *S. Cerevisiae*, *S. Octosporus*, *S. Pombe*, *Candida albicans*, *Candida glabrata*,*Candida tropicalis*,*Candida dubliniensis*,*Candida parapsilosis*. The number of genes of this network ranges from 8 in *S. Cerevisiae*, *S. Pombe*, *S. Japonicus* and *S. Octosporus* to 4 in the candida species. The tree shows ticker lines for worst match with respect the tree of Figure 2 (RAM network, case study 3), with overall topological score of 61.5, see [28] for details. To explore the capability of our approach in comparison to other regression-based motif discovery methods, we applied MatrixREDUCE [12] to the same sets of sequence and expression data. MatrixREDUCE also exploits the correlation between gene expression levels and the occurrence frequency of short DNA segments in upstream sequences to discover binding motifs, but differs from our framework in two following points. Firstly, it employs a deterministic forward variable selection scheme. Motifs are added to the regression model looking at their matching coefficients in descending order. Although this approach is attractive by its computational simplicity, it is more likely to get trapped in local maxima than the stochastic sampling process employed by our procedure. Secondly, our proposed approach provides a principled way to account for multiple experiments/replicates simultaneously while MatrixREDUCE is applicable to one experiment condition at a time. From the results in table 1, MatrixREDUCE could detect only parts of the consensus motif sequences in *S. Pombe* and *S. Japonicus*. We show in the supplementary material the motif marginal probabilities for $m_{\text{prior}} = 1$ of *S. Pombe* (Figure S1a in the supplementary material), *S. Japonicus* (Figure S1b in the supplementary material), *S. Octosporus* (Figure S1c in the supplementary material).

Case study 2: cytokinesis transcriptional network in *Candida* clade

In the second case study we consider the cytokinesis process of *Candida albicans* in the context of three other closely-related species *Candida dublininensis*, *Candida tropicalis*, and *Candida parapsilosis*. While the *Candida* clade is in close approximation to the fission yeast *S. pombe* in the phylogenetic tree, it does not go through the genome duplication event.

The cytokinesis and mother-daughter cell separation processes of *C. albicans* have been associated with 94 genes, whose expression levels periodically peak during the G2/M phase over four biological replicates [29]. Using this gene set as the starting point, we applied hierarchical clustering with Bayesian similarity measurement [30] on their expression profiles to separate regulatory networks with different transcription factors. Eight

Table 1. Motifs detected for septation transcriptional network in fission yeast clade (case study 1).

Motifs detected in *S. Pombe*			
by Bayesian variable selection	**marginal probability**	**by MatrixREDUCE**	**t-value**
GGT **GGCTGG**CA	0.995872	**CCAG**	−8.495
AATGTAA	0.992299833	AT	7.854
GTGGTTGG	0.990841403	TCTG	−7.146
TTGCTTTAT	0.966439608	GAGAA	−6.868
GAAAATCGAA	0.964823733	CCTC	−6.416
ATCGATGGTAA	0.964302733	TCCTC	−6.119
CAAGAAAGTAC	0.952851275	CG	−6.017
TCAATAT **CCAGC**	0.930580914	AT	5.857
GATTTTACCA	0.930109523	CTCT	−5.723
TTAT **CCAGCC**	0.913970665	TTC	−5.602
GTAAAAAA	0.911840485		
AAATTTAAGAG	0.899786547		
TTATATAA	0.892315745		
CAAATATAAA	0.8920356		
CATGGCGGG	0.868056013		
TCTATATTCGG	0.773856698		
TTACTTTCTT	0.728913078		
Motifs detected in *S. Japonicus*			
by Bayesian variable selection	**marginal probability**	**by MatrixREDUCE**	**t-value**
ACTCGCGTCAC	0.962531425	A **CAGC**G	−15.946
AAGGA **GGCT**	0.88789641	GCAT	8.422
ACGGTGTGAA	0.860245247	TCGGT	−7.192
GGCTGG	0.8282509	TCGGT	−5.905
A **GGCTGG**T	0.789977544	TTTTCC	−5.707
CAGATTTCGTGC	0.777363805	TCGGT	−5.457
A **CCAGCC**	0.742744755	TTTTTT	−5.412
GTGTCAC	0.72705333	AGGA	5.361
ATGCATA	0.70449432	TTTTT	−5.351
		CTAA	5.292
Motifs detected in *S. Octosporus*			
by Bayesian variable selection	**marginal probability**	**by MatrixREDUCE**	**t-value**
GAT **GGCTGG**TA	1	CG	−8.395
GTATCGGTTG	0.998044545	TCGAA	−7.167
GTTGCAAGT	0.997534734	CTTGA	7.12
TTGTTTGTTTA	0.99644855	AGACA	−7.12
ACTTTCATCCA	0.99155467	AGATTT	−6.574
ACAATGGAT	0.98432507	AGGC	−6.574
CCTTCCACCGA	0.980347067	AAGAT	−6.17
TCAAT **CCAG**T	0.975195473	GATCA	−5.87
TTCGTTTCCGT	0.955906395	ACTGAA	−5.287
CATTCAGGGG	0.955746803	AAGATTT	−5.287
ACTTTACTC	0.92077283		
AGAGAGAAA	0.91857454		
GGTACGAAGAA	0.911070636		
AAGAGCAGAGC	0.88196748		
CGTCGTGGTG	0.870479665		
GTTCGATGGC	0.83167287		

Table 1. Cont.

Motifs detected in *S. Octosporus*			
by Bayesian variable selection	**marginal probability**	**by MatrixREDUCE**	**t-value**
GATTTTACTCG	0.7601617		
GTAGAAACA	0.757162374		

doi:10.1371/journal.pone.0042489.t001

tightly co-regulated genes were chosen for our further analysis: Cdc5, Chs1, Hof1, Kip2, Chs8, Fgr29, and two less known genes orf19.1334 and orf19.6119. Among the selected genes, the first five are known to preserve their functions as compared to *S. cerevisiae*, and the latter three are specific to *C. albicans* only, without any orthologs in the other two model budding and fission yeast.

We applied Bayesian variable selection and MatrixREDUCE to detect binding motifs in the four *Candida* species using time-series microarray expression data of *C. albicans* and upstream sequences from all four species. The expression data were collected from four independent biological replicates [29]. The upstream sequences for *C. Albicans*, *C. Tropicalis*, and *C. Parapsilosis* were obtained from the MIT Broad Candida database [26] , and *C. Dubliniensis* from Sanger Institute Sequencing Project [31]. We show a summary of the results obtained using all four experiments for sparsity setting $m_{\text{prior}} = 5$ in Table 2. While motifs with longer width are treated more favorably, the algorithm is able to pick up a number of patterns. Two emergent patterns that occur repeatedly in all four

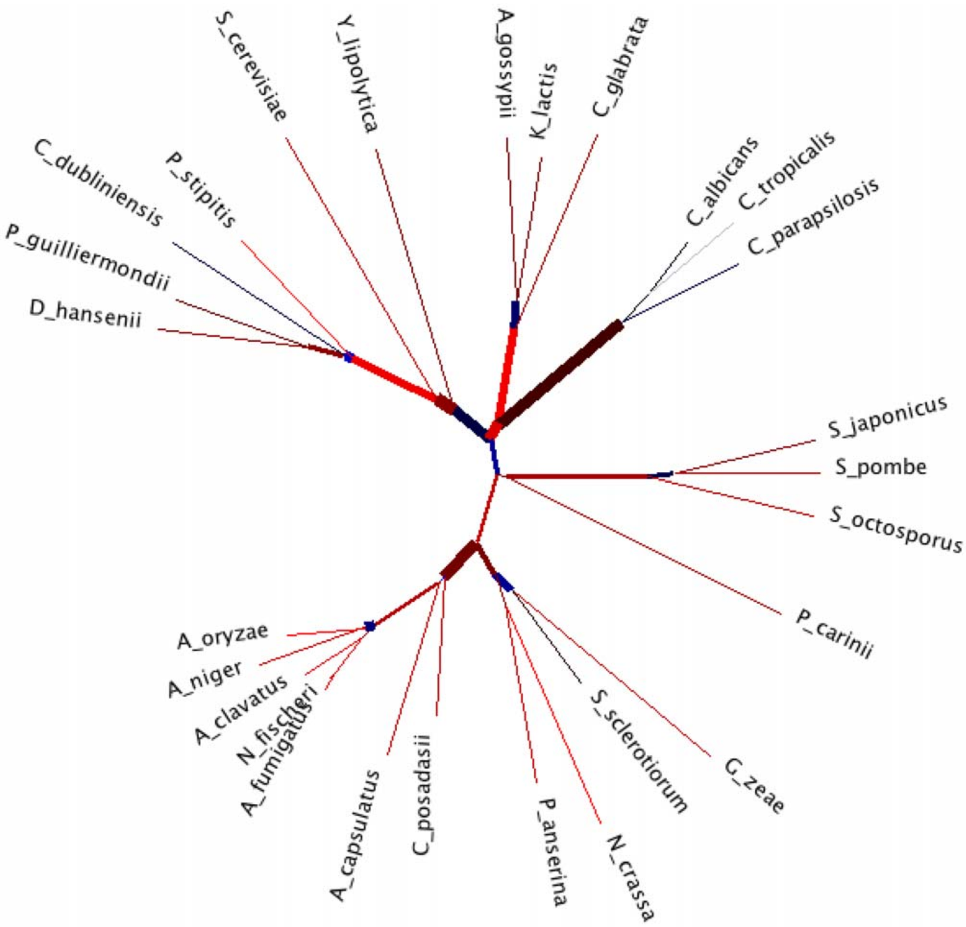

Figure 1. ENG1 ML tree. Maximum Likelihood tree, based on JTT model of evolution, inferred using Eng1 protein sequence from the following species: *S. Japonicus*, *S. Octosporus*, *S. Cerevisiae*, *S. Pombe*, *Kluyveromyces lactis*, *Debaryomyces hansenii*, *Candida Albicans*, *Yarrowia lipolytica*, *Aspergillus oryzae*, *Phaeosphaeria nodorum*, *Neurospora crassa*, *Vanderwaltozyma polyspora Neosartorya fischeri Pichia guilliermondii*,*Coccidioides posadasii*, *Gibberella zeae*, *Ashbya gossypii*, *Sclerotinia sclerotiorum*, *Magnaporthe grisea*, *Ajellomyces capsulatus*, *Aspergillus clavatus*, *Aspergillus niger*, *Pichia stipitis*, *Lodderomyces elongisporus*, *Candida glabrata*,*Candida Tropicalis*,*Candida dubliniensis*,*Candida parapsilosis*; *Brassica napus* and *Sorangium cellulosum* are plant sequences used as outgroups, i.e. to facilitate the rooting of fungi phylogeny; we also include *S. Japonicus* Eng1 and Eng2 proteins and *S. Cerevisiae* Acf1 and Acf2 proteins. From a methodological purpose, we validate this phylogeny with a phylogeny with the same number of species, based on cdc5, a regulator of G2/M transition of mitotic cell cycle with the same visualisation as in [28]; the width corresponds to phylogenetic agreement.
doi:10.1371/journal.pone.0042489.g001

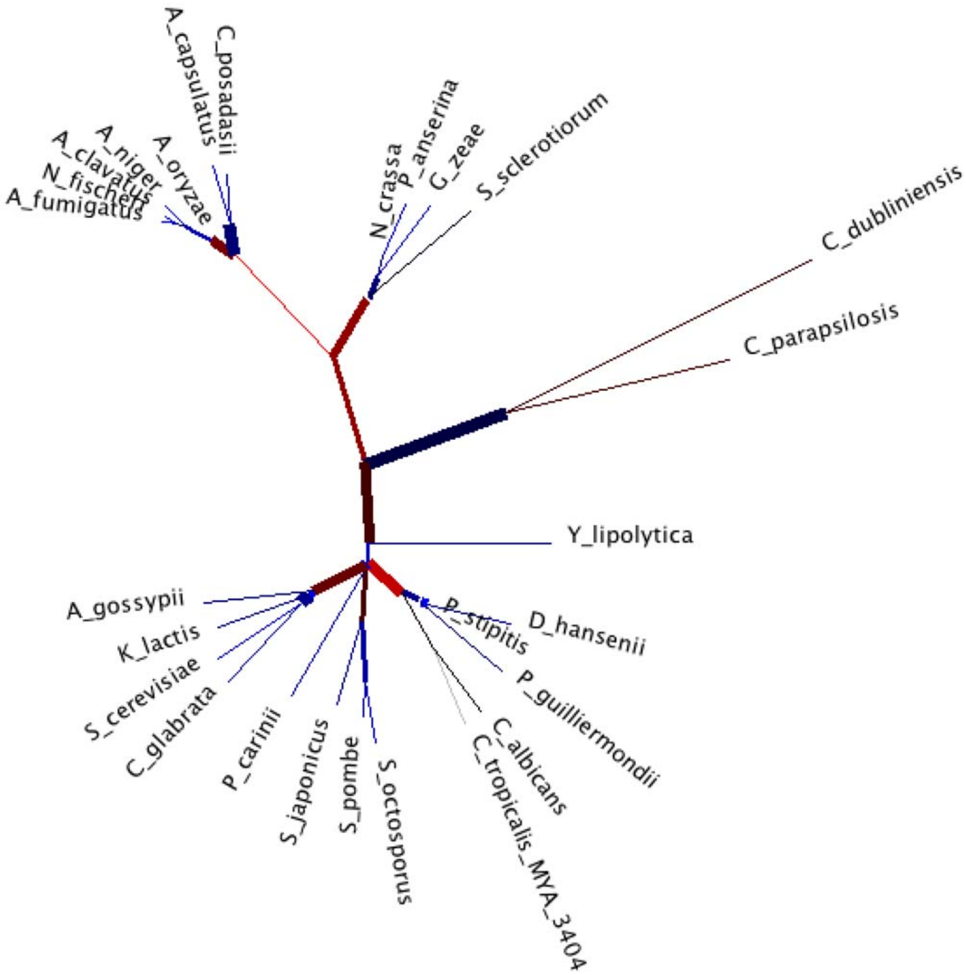

Figure 2. RAM ML tree. Maximum Likelihood tree, based on JTT model of evolution, inferred using RAM protein sequences from the species of figure 1. We validate this phylogeny with a phylogeny with the same number of species, based on cdc5, a regulator of G2/M transition of mitotic cell cycle with the same visualisation as in [28].
doi:10.1371/journal.pone.0042489.g002

species are TCATTC and TCAATT (which were printed bold in the tables for easier comparison). These are suggestively the variants of the consensus motif TCA(A/T)T(C/T). The results from MatrixREDUCE are also presented in the same table. In agreement with the discussion of method comparison in the first case study, Bayesian variable selection definitely adds significant value from its comprehensive combinatorial effect search. While MatrixREDUCE could identify some parts of the proposed motif in *C.* Albicans, it fails to do so for the other related species. Note that *C. Parapsilosis* has been observed in clinical literature [32,33] to have distinct features in comparison to *C. albicans*, *C. Dubliniensis*, and *C. Tropicalis*. While the other three species are strictly human pathogen, *C. Parapsilosis* is also found in a wide range of environments including animals, soils, and the sea. Such flexibility might suggest corresponding shift in its cell cycle regulatory mechanism. We show in the supplementary material the motif marginal probabilities for the following species: *C. Albicans* (Figure S2a: $m_{\mathrm{prior}}=1$; Figure S2b: $m_{\mathrm{prior}}=3$; Figure S2c: $m_{\mathrm{prior}}=5$), *C. Dubliniensis* (Figure S3a $m_{\mathrm{prior}}=1$; Figure S3b $m_{\mathrm{prior}}=3$; Figure S3c $m_{\mathrm{prior}}=5$), *C. Tropicalis* (Figure S4a $m_{\mathrm{prior}}=1$; Figure S4b $m_{\mathrm{prior}}=3$; Figure S4c $m_{\mathrm{prior}}=5$), *C. Parapsilosis* (Figure S5a $m_{\mathrm{prior}}=1$; Figure S5b $m_{\mathrm{prior}}=3$; Figure S5c $m_{\mathrm{prior}}=5$).

Case study 3: RAM transcriptional network in the *Ascomycota*

RAM (regulation of Ace2p transcription factor and polarized morphogenesis) is a conserved signaling network that regulates polarized morphogenesis in yeast, worms, flies, and humans [34]. In unicellular fungi, the RAM network comprises the proteins Cbk1, Mob2, Kic1, Hym1, Sog2, and Tao3. *S. Cerevisiae* strains harboring mutations in any of these genes display cell separation defects and a loss of polarity. The ability of *C. Albicans* to undergo morphogenesis from yeast to hyphal form is associated to the condition of causing the disease; the RAM genes CaCBK1, CaMOB2, CaKIC1, CaPAG1, CaHYM1, and CaSOG2 are good candidates for drugs controlling the growth of the organism and therefore the spreading of the infection [35]. Figure 2 shows the maximum likelihood phylogenetic tree obtained using CaCBK1 gene sequences from the same species of Figure 1.

Microarray data of *C. Albicans* were obtained from a recent investigation of RAM network's role in cell polarity and hyphal morphogenesis processes [35]. Four single-replicate experiments were conducted on wild-type (SC5314) and CaMOB2 mutant strains grown in either normal yeast or hypha-inducing serum medium. The identified RAM-dependent hypha-specific genes

Table 2. Motifs detected for cytokinesis transcriptional network in Candida clade (case study 2).

Motifs detected in *C. Albicans*

by Bayesian variable selection	marginal probability	by MatrixREDUCE	t-value
T **TCATTC**ATTC	0.978005	AATGAA	10.56
TCAATT	0.900631	AT **AATT**	−8.362
TTGAT	0.666508	AAATGAA	8.143
TGAAATCA	0.663159	TGAAAT	7.691
ATGAAATA	0.650006	**CAAT**	7.129
GAAACTGA **AATT**	0.637142	AAGTT	6.78
AATTAATT	0.626425	AATGAAT	6.09
		GTTGTTG	6.09
		ATGAA	6.086
		AATTAAT	6.086

Motifs detected in *S. Dubliniensis*

by Bayesian variable selection	marginal probability	by MatrixREDUCE	t-value
GTT **AATT**CCA	0.999259	ATTT	8.486
AGTTCA	0.962802	CAACAA	8.067
TTTCCTGATTTG	0.925764	CATCA	−7.891
CCTA **AATT**AAG	0.870212	GTCT	7.473
AAT **TCAATT**	0.82158	ACAACAAC	7.471
TCCTGA	0.804322	CAAAATA	7.192
TATGCAA	0.68007	AGG	−7.101
TCATTCCACTT	0.58234	CAACAACA	6.904
TCAAT	0.56146	CAACAAC	6.058
TCCTGATTTG	0.560856	ACAACAA	5.94
TTCGTC	0.548985		

Motifs detected in *S. Tropicalis*

by Bayesian variable selection	marginal probability	by MatrixREDUCE	t-value
TAATG **CATT**	0.74355	CCATG	12.846
AATTT	0.699152	TATTTAT	11.433
TAATGAAA	0.684862	TTATTTAT	10.903
TGAAACT**TTGA**A	0.662662	TTATTTA	10.267
ATTTGGTCA	0.562592	TTTATTTA	9.598
		TATTTATT	8.848
		ATTTAT	8.701
		ATTTATTT	8.645
		TAACA	8.432
		GTTGGT	−8.432

Motifs detected in *S. Parapsilosis*

by Bayesian variable selection	marginal probability	by MatrixREDUCE	t-value
AT **TCAAT**	0.671743	AGAGA	10.182
GC **CAATT**C	0.548763	CCAT	−9.971
AGATAAGCA	0.547218	AGAG	7.773
TTCCA **AATT**	0.54143	GAGAT	7.626
		GAGA	7.356
		AGAGAG	7.26
		AAAAC	−6.469
		CAAACA	−6.384
		AAAC	−5.884
		CTA	−5.504

doi:10.1371/journal.pone.0042489.t002

Table 3. Motifs detected for RAM transcriptional network in the Ascomycota (case study 3).

Motifs detected in ***C. Albicans***			
by Bayesian variable selection	**marginal probability**	**by MatrixREDUCE**	**t-value**
AATGAG **AAATA**A	1	CCAA	10.494
GAGTTGA	0.7266	CCATA	−10.281
GAGA **AAAGA**AAA	0.7263	GATTAC	8.597
A **AAAT**	0.6174	ATCAC	8.591
ACTTT **TCTT**A	0.4965	CTAAA	8.591
TATTT	0.4731	**TATT**GA	−8.383
TGGATTTTG	0.4458	TCATAT	−7.47
AGAC **AAGA**	0.4028	ACTCT	7.47
AAAATGAA	0.3884	AGGC	7.104
CTTT **TCTT**	0.3768	A **ATTT**	−6.759
TTGAC **CTTT**	0.3666		
T **TATT**	0.3254		
TATTGGA	0.319		
Motifs detected in ***S. Dubliniensis***			
by Bayesian variable selection	**marginal probability**	**by MatrixREDUCE**	**t-value**
AAATGAAAGG	0.982	CTCT	19.105
AAATTCAATTTC	0.7966	ATGTTT	−8.855
GAAAA	0.7782	AGGTG	8.597
AAAAC	0.7464	GGGT	8.431
ATTTC **TATTT**	0.7382	G **TATT**	8.297
TCAGTTTTAA	0.6154	ATGTT	−8.022
TTA **CTTT**TCT	0.6129	TAGG	7.959
TGGTAG **TATT**G	0.3549	TCTCTC	7.945
CTTTT **TCTT**	0.3287	GGTC	7.104
TATATATATATA	0.3155	CTCTC	6.883
Motifs detected in ***S. Tropicalis***			
by Bayesian variable selection	**marginal probability**	**by MatrixREDUCE**	**t-value**
AATTAA **TATT**CG	0.9516	C **AATA**	−15.527
GTCTGT	0.8571	TCCAA	10.351
TCTGTAGATAG	0.553	A **AAAT**G	9.408
TACACGAACAAT	0.4903	CAATTA	−8.855
GTAGAGG **AAGA**	0.4588	TC **AATA**	−8.855
CTGTAG	0.3561	TGCAA	8.383
TACACGC	0.2928	TTGC	7.979
ATTGT	0.2789	GTATA	7.959
AAAGAGACAC	0.2734	CCAAG	−7.47
ACAAC	0.2242	GGATT	7.47
GACAA	0.2179		
AGTGT	0.2151		
Motifs detected in ***S. Parapsilosis***			
by Bayesian variable selection	**marginal probability**	**by MatrixREDUCE**	**t-value**
TGTGGTGGT	0.8894	ATGTG	−12.033
ACATACACCCC	0.5923	GAA	11.016
CAC **AAGA**GCAA	0.3663	TCGC	−10.351
AGTTGATCCTA	0.3434	ACAAAC	10.351
AGTTG	0.3056	GATGT	−10.281

Table 3. Cont.

Motifs detected in *S. Parapsilosis*			
by Bayesian variable selection	**marginal probability**	**by MatrixREDUCE**	**t-value**
GAGAACCGTT	0.297	TGAAC	8.597
TACAACCC	0.2714	**AAAT**GA	8.597
ACTTG	0.2643	CAAGT	8.383
GGTATGTGTGTA	0.2484	G **AAGA**A	8.112
A **CTTT**AA	0.205	ATGCA	−7.47

doi:10.1371/journal.pone.0042489.t003

suggested the association of RAM network with Tup1p/Nrg1p-regulated morphological processes.

A summary of the results for all four species is shown in Table 3. The left two columns list out motifs detected by Bayesian variable selection with their marginal probabilities averaged over 8 control subsets. The right two columns list out motifs detected by MatrixREDUCE with corresponding T-values. To account for the possible depreciation of motif effects in various species, we lowered the cutting threshold of marginal effect to 0.3 for *C. Albicans* and *C. Dubliniensis*, and 0.2 for *C. Tropicalis* and *C. Parapsilosis*. Two emergent patterns that occur repeatedly in all four species are AAAGA and AAATA, which constitute the consensus motif AAA(G/T)A. Again the comparison with MatrixREDUCE results suggest the capability of our approach to detect more comprehensive patterns. We present the motif marginal probabilities of *C. Tropicalis* for $m_{\text{prior}}=1$ (Figure S6a of the supplementary material); for $m_{\text{prior}}=3$ (Figure S6b of the supplementary material) and for $m_{\text{prior}}=5$ (Figure S6c of the supplementary material).

Further points of discussion and Conclusions

Here we present a general algorithm to predict sequence motifs in a newly sequenced species combining linear regression, Bayesian variable selection and experimental data of a well characterized model organism. First we apply our method within the same species (with several replicas), then we extend it to phylogenetically close model species. The proposed method for regulatory motif discovery do not rely on previous knowledge of co-regulated sets of genes, and in that way differ from the main stream literature on computational motif discovery. The validation of the discovered motifs relies on previously published regulatory motifs in yeasts and on an internal testing procedure. Finally, we apply it to make prediction on three important eukaryotic gene networks and compare the results with a currently used method. We believe that the tables of regulatory sequences we present could be useful to genome researchers because the motifs represent putative regulatory sites and commonalities among the studied species.

Most of the species we have considered are recently sequenced with little annotations available. We have searched the proposed sets of motifs in a number of available low eukaryotes genomic resources such as NCBI , http://www.pombase.org/, http://www.broadinstitute.org/annotation/genome/ and others. We found that motifs with the highest marginal probability have been reported in low eukaryotes and plants. For example GTTAATTCCA (motif detected in *S. Dubliniensis*) has been reported being a target sequence in the phenobank in *C. Elegans* database [36]. The motif TCAATCCAGT (found in *S. Octosporus*) occurs in the promoter region of the locus AT5TE39210 of *A. Thaliana*; ACAATGGAT (found in *S. Octosporus*) is conserved in the promoter of several plants such as barley and wheat (CA679037, homologous to DATFAP); GTATCGGTTG (found in *S. Octosporus*) is a motif reported (http://yeastract.com/) in the promoter of YBR242w, which is a protein of unknown function of *S. Cerevisiae* that localizes to the cytoplasm and nucleus. The motif ATCGATGGTAA (found in *S. Pombe*) has been reported to regulate the expression of the GLN1 in *S. Cerevisiae* [37]; ACTTTCATCCA (found in *C. Albicans*) is reported in regulatory elements in promoters of defense genes (GRX480) in *A. Thaliana* [38].

We also find that some of the motifs ((for example TTTCCTGATTTG and AATGAGAAATAA) have been described in databases automatically built by a number of computational tools such as http://www.cisred.org/ [39], ABS, http://genome.crg.es/datasets/abs2005/ [40] and Phylonet (http://stormo.wustl.edu/molee//Motif/). Some motifs produced hits in high eukaryotes or bacteria sequences (not reported because the species are too distant from the fungi); short motifs produced many hits.

We believe that the conservation of the size of the network across the species and its functional role are very important. Indeed, the first two examples, which are both cell-cycle related with relatively conserved gene network size across the species, give a much better result than the third case (RAM network). Given the high cost of performing a large number of experimental replicates, we make the hypothesis that experimental evidences, from species similar to that under analysis, may provide additional statistical support. We can assume that the closer the species to the one under investigation, the better is. The most interesting result presented in the paper is to show that the marginal probabilities become much higher (about 3 fold) than those obtained using single replicates and one species [8]. It is interesting to consider our work in the light of a recent discussion about experimental design in the context of phylogenetic inference [41]. A first simple question is whether the number of genes involved in the transcriptional network is the same for the different species considered or whether it varies when the phylogenetic distance increases. This is an important point raised in [42], where the authors have demonstrated an inverse correlation between the rate of evolution of transcription factors and the number of genes they regulate. For small gene networks, distant species may not provide adequate support and, in general, the distance may depend on the size of the genetic network. Noteworthy, while much effort has been focused on sequencing the genomes of widely divergent species, recently there has also been interest in sequencing the genomes of closely related species, with the target of comparing and contrasting them for subtle differences [41]. There exist different papers to quantify the utility of multiple genomes for the detection of conserved DNA regions ([41], [43], [44], [45]). Margulies et al. [46] described an economically efficient approach to show that low redundancy sequencing of additional genomes is a useful first step in locating conserved regions in the species of interest. Current efforts in sequencing may allow to sequence 'on demand' in order to shed light on an important regulatory network under study. We believe that our statistical approach can result of some practical utility for outputting putative regulatory sites and annotating new genomes. We have also found that it highlights interesting statistical problems raised by high throughput data integration, such as sequence and gene expression. We believe that this methodology could be extended to integrate other high-throughput omic data.

Supporting Information

Figure S1 Motif marginal probabilities, case study 1. Posterior marginal probabilities of (a) *S. Pombe* , (b) *S. Japonicus*, (c) *S. Octosporus* candidate motifs for $m_{\text{prior}}=1$.
(TIFF)

Figure S2 *C. Albicans* motif marginal probabilities, case study 2. Posterior marginal probabilities of *Candida Albicans* candidate motifs for (a) $m_{\text{prior}}=1$; (b) $m_{\text{prior}}=3$; (c) $m_{\text{prior}}=5$.
(TIFF)

Figure S3 *C. Dubliniensis* motif marginal probabilities, case study 2. Posterior marginal probabilities of *Candida Dubliniensis* candidate motifs for (a) $m_{\text{prior}}=1$; (b) $m_{\text{prior}}=3$; (c) $m_{\text{prior}}=5$.
(TIFF)

Figure S4 *C. Tropicalis* motif marginal probabilities, case study 2. Posterior marginal probabilities of *Candida Tropicalis* candidate motifs for (a) $m_{\text{prior}}=1$; (b) $m_{\text{prior}}=3$; (c) $m_{\text{prior}}=5$.
(TIFF)

Figure S5 *C. Parapsilosis* motif marginal probabilities, case study 2. Posterior marginal probabilities of *Candida Parapsilosis* candidate motifs for (a) $m_{\text{prior}}=1$; (b) $m_{\text{prior}}=3$; (c) $m_{\text{prior}}=5$.
(TIFF)

Figure S6 *C. Tropicalis* motif marginal probabilities, case study 3. Posterior marginal probabilities of *Candida Tropicalis* candidate motifs for (a) $m_{\text{prior}}=1$; (b) $m_{\text{prior}}=3$; (c) $m_{\text{prior}}=5$.
(TIFF)

Acknowledgments

We acknowledge the *Schizosaccharomyces Japonicus* Sequencing Project at Broad Institute of Harvard and MIT (http://www.broad.mit.edu). Claudia Angelini and Italia De Feis acknowledge the Italian Flagships Project "Epigenomics", that partially supported the research.

Author Contributions

Conceived and designed the experiments: PL CA ID VN. Performed the experiments: PL CA ID VN. Analyzed the data: PL CA ID VN. Contributed reagents/materials/analysis tools: PL CA ID VN. Wrote the paper: PL CA ID VN.

References

1. Tompa M, Li N, Bailey TL, Church GM, De Moor B, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23: 137–144.
2. Brown CT (2008) Computational approaches to finding and analyzing cis-regulatory elements. Methods Cell Biol 87: 337–65.
3. Zheng W, Zhao H, Mancera E, Steinmetz LM, Snyder M (2010) Genetic analysis of variation in transcription factor binding in yeast. Nature 464:1187–1191.
4. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497–1502.
5. Weirauch MT, Hughes TR (2010) Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. Trends Genet 26:66–74.
6. Conlon EM, Liu XS, Lieb JD, Liu JS (2003) Integrating regulatory motif discovery and genome-wide expression analysis. Proc Natl Acad Sci USA 100: 3339–3344.
7. Liu XS, Brutlag DL, Liu JS (2002) An algorithm for finding protein-dna binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat Biotechnol 20: 835–839.
8. Angelini C, Cutillo L, De Feis I, van der Wath R, Liò P (2007) Identifying regulatory sites using neighborhood species. Lecture Notes in Computer Science 4447: 1–10.
9. Tadesse MG, Vannucci M, Liò P (2004) Identification of dna regulatory motifs using Bayesian variable selection. Bioinformatics 20: 2553–2561.
10. Angelini C, Cutillo L, De Feis I, Nguyen VA, van der Wath R, et al. (2010) Combining Replicates and Nearby Species Data: A Bayesian Approach. Computational Intelligence Methods for Bioinformatics and Biostatistics. Lecture Notes in Computer Science 6160: 367–381.
11. Conlon EM, Song JJ, Liu JS (2006) Bayesian models for pooling microarray studies with multiple sources of replications. BMC Bioinformatics 7: 247–250.
12. Foat BC, Houshmandi SS, Olivas WM, Bussemaker HJ (2005) Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. PNAS 102: 17675–17680.
13. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8: 275–282.
14. Wolfe KH, Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387: 708–713.
15. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium, Nature Genet 25: 25–29.
16. Wapinski I, Pfeffer A, Friedman N, Regev A (2007) A natural history and evolutionary principles of gene duplication in fungi. Nature 449: 54–61.
17. Wingender E, Dietze P, Karas H, Knüppel R (1996) TRANSFAC: A Database on Transcription Factors and Their DNA Binding Sites. Nucl Acids Res 24: 238–241.
18. Smit A, Hubley R, Green P. RepeatMasker. Available: http://repeatmasker.org.
19. Angelini C, Cutillo L, De Feis I, Liò P, van der Wath R (2008) Combining experimental evidences from replicates and nearby species data for annotating novel genomes. AIP proceedings 1028: 277–291.
20. Martín-Cuadrado AB, Dueñas E, Sipiczki M, Vázquez de Aldana CR, del Rey F (2003) The endo-β-1,3-glucanase eng1p is required for dissolution of the primary septum during cell separation in Schizosaccharomyces Pombe. J Cell Sci 116: 1689–1698.
21. Dekker N, Speijer D, Grun CH, van den Berg M, de Haan A, Hochstenbach F (2006) Role of the alpha-glucanase Agn1p in fission-yeast cell separation. Mol Biol Cell 15: 3903–3914.
22. Rustici G, Mata J, Kivinen K, Liò P, Penkett CJ (2004) Periodic gene expression program of the fission yeast cell cycle. Nature Genet 36: 809–817.
23. Mulhern SM, Logue ME, Butler G (2006) Candida albicans Transcription Factor Ace2 Regulates Metabolism and Is Required for Filamentation in Hypoxic Conditions. EUKARYOTIC CELL 5: 2001–2013.
24. Martín-Cuadrado AB, Encinar del Dedo J, de Medina-Redondo M, Fontaine T, del Rey F, et al. (2008) The Schizosaccharomyces Pombe endo-1,3-β-glucanase Eng1 contains a novel carbohydrate binding module required for septum localization. Molecular Microbiology 69: 188–200.
25. Oliva A, Rosebrock A, Ferrezuelo F, Pyne S, Chen H, et al. (2005) The cell cycle regulated genes of schizosaccharomyces pombe. PLoS Biol 33: 1239–1260.
26. Broad MIT Candida and Schizosaccharomyces database. Available: http://www.broadinstitute.org/science/data.
27. Piazza F, Liò P (2005) Statistical analysis of low–complexity sequences in the human genome. Physica A 347: 472–488.
28. Nye TM, Liò P, Gilks WR (2006) A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. Bioinformatics 22:117–119.
29. Cote P, Hogues H, Whiteway M (2009) Transcriptional analysis of the Candida albicans cell cycle. Molecular Biology of the Cell 20: 3363–3373.
30. Nguyen VA, Liò P (2009) Measuring similarity between gene expression profiles: a Bayesian approach. BMC Genomics 10 suppl. 3: S14.
31. Candida dubliniensis sequencing project. Available: http://www.sanger.ac.uk/sequencing/Candida/dubliniensis/.
32. Weems JJ Jr (1992) Candida parapsilosis: epidemiology, pathogenicity, clinical manifestations, and antimicrobial susceptibility. Clin Infect Dis 14, 3: 756–66.
33. Trofa D, Gacser A, Nosanchuk J (2008) Candida parapsilosis, an Emerging Fungal Pathogen, Clin Microbiol Rev21, 4: 606–625.

34. Nelson B, Kurischko C, Horecka J, Mody M, Nair P, et al. (2003) RAM: a conserved signaling network that regulates Ace2p transcriptional activity and polarized morphogenesis. Mol Biol Cell 14: 3782–3803.
35. Song Y, Cheon SA, Lee KE, Lee SY, Lee BK, et al. (2008) Role of the RAM network in cell polarity and hyphal morphogenesis in Candida albicans. Mol Biol Cell 19:5456–77.
36. Sönnichsen B, Koski LB, Walsh A, Marschall P, Neumann B, et al. (2005) Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans. Nature 434: 462–469.
37. Minehart PL, Magasanik B (1992) Sequence of the GLN1 Gene of Saccharomyces cerevisiae: Role of the Upstream Region in Regulation of Glutamine Synthetase Expression. Journal of Bacteriology 174: 1828–1836.
38. Ndamukong I (2006) Characterization Of An Arabidopsis Glutaredoxin that Interacts With Core Components Of The Salicylic Acid Signal Transduction Pathway - Its Role In Regulating the Jasmonic Acid Pathway PhD Thesis Gottingen 28-02-2006 (http://webdoc.sub.gwdg.de/diss/2006/ndamukong/ndamukong.pdf).
39. Robertson AG, Bilenky M, Lin K, He A, Yuen W, et al. (2005) cisRED: A database system for genome scale computational discovery of regulatory elements. Nucleic Acids Res (Database issue) 1, 34:D68–73.
40. Blanco E, Farre D, Alba M, Messeguer X, Guigo R (2006) ABS: a database of Annotated regulatory Binding Sites from orthologous promoters. Nucleic Acids Research 34:D63–D67.
41. Geuten K, Massingham T, Darius P, Smets E, Goldman N (2007) Experimental design criteria in phylogenetics: where to add taxa. Syst Biol 56: 609–622.
42. Rajewsky N, Socci ND, Zapotocky M, Siggia ED (2002) The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons. Genome Res 12: 298–308.
43. McAuliffe JD, Jordan MI, Pachter L (2005) Subtree power analysis and species selection for comparative genomics. Proc Natl Acad Sci USA 102: 7900–7905.
44. Eddy SR (2005). A Model of the Statistical Power of Comparative Genome Sequence Analysis. PLoS Biol 3(1): e10. doi:10.1371/journal.pbio.0030010.
45. Pardi F, Goldman N (2005) Species choice for comparative genomics: being greedy works. PLoS Genet 1(6):e71.
46. Margulies EH, Vinson JP, Miller W, Jaffe DB, LindbladToh K (2005) An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. Proc Natl Acad Sci USA 102: 4795–4800.